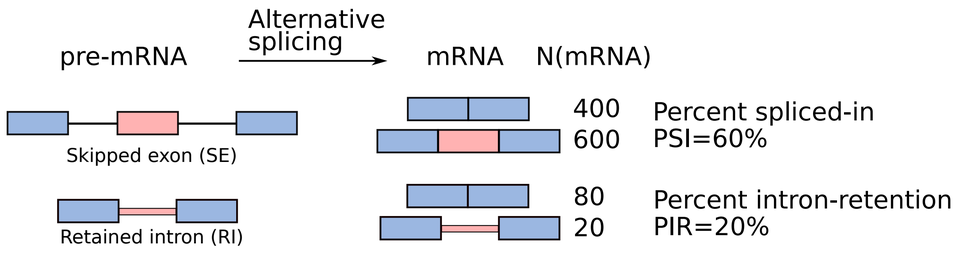
NGS 2017
Barcelona

Matt: a command-line toolkit for the downstream analysis of splicing events

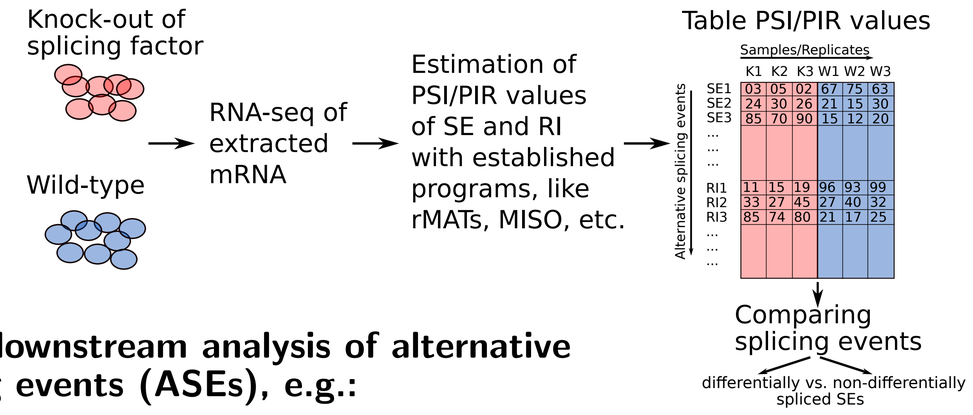
André Gohr[1,2], Manuel Irimia[1,2]

agohr.crg.eu/matt

Alternative splicing



Analysis of alternative splicing events



Goal: downstream analysis of alternative splicing events (ASEs), e.g.:

1. Automatic extraction of features of exons/introns
2. Detection of features discriminating sets of exons/introns
3. Predicting hits of binding motifs of RNA-binding proteins (RBPs)
4. Enrichment analysis with hits of RBP binding motifs
5. Plotting positional distributions of predicted hits of binding motifs

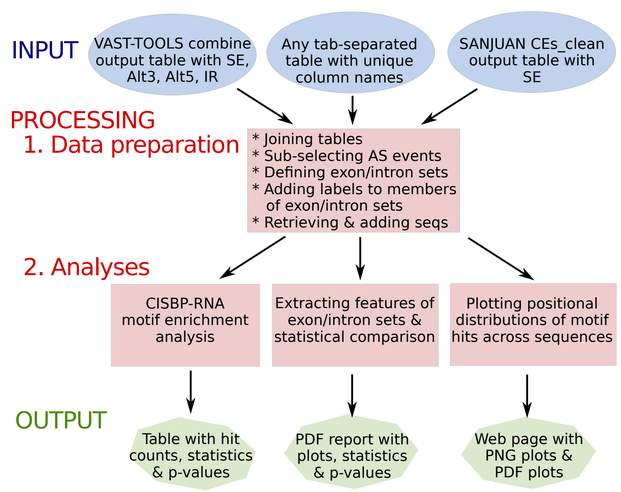
Problem: comprehensive tools missing, complicating the investigation of splicing mechanisms and alternative splicing

Solution: toolkit Matt for downstream analysis of splicing events

Structure of input/output data and workflow

Tab-separated, text **tables** with unique column names. **Advantages**:

1. **Compatibility**: programs for estimation of PSI/PIR values often report results in tables
2. **User-friendliness**: tables can be easily studied with any spread sheet program
3. **Clarity**: column names document type of data in columns
4. **Synergy**: Matt commands (but also Awk, Sed, etc.) can be combined using piping



Automatic extraction of 75/50 features of exons/introns

* Regions considered for feature extraction:

* Extracted features include:

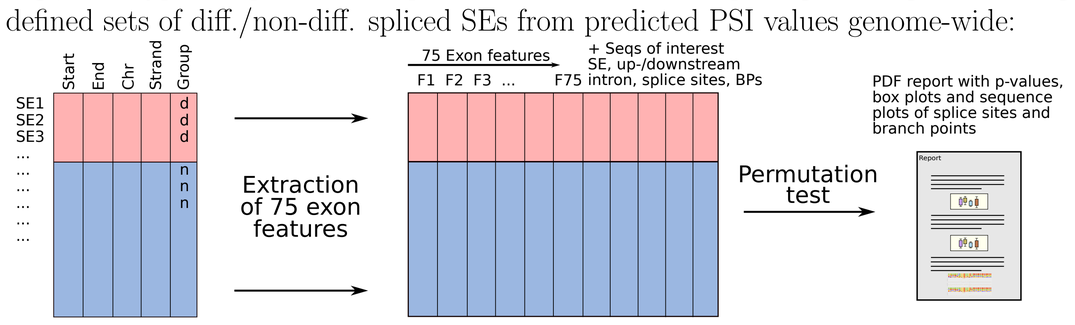
* Additional features:

- Ratios of exon/exon lengths and GC content, and exon/intron lengths and GC content
- position of exon/intron within transcript
- Number of transcripts, median transcript length
- Exon-exon and intron-intron co-occurrences

Matt invokes maximum-entropy models for computation of splice site strength ([1]) and SVM-BPfinder for extraction of branch point features ([2]).

Detecting features discriminating sets of exons/introns

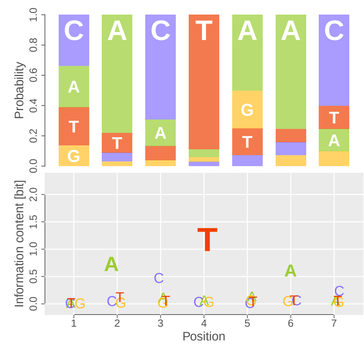
E.g., wild-type samples vs. samples with a knocked-out splicing-factor gene: Having defined sets of diff./non-diff. spliced SEs from predicted PSI values genome-wide:



Binding motifs in Matt

Perl regular expressions defining complex consensus sequences, e.g., the Nova YCAY cluster motif `[CT]CA[CT](.{0,3}[CT]{0,1}CA[CT]|.{4,23}[CT]CA[CT]){2}`, and position weight matrices (PWM), e.g., human SF1

SYM	POS_1	POS_2	POS_3	POS_4	POS_5	POS_6	POS_7
A	0.272	0.780	0.174	0.051	0.502	0.755	0.148
C	0.338	0.058	0.693	0.029	0.072	0.086	0.601
G	0.137	0.031	0.039	0.030	0.249	0.072	0.099
T	0.252	0.131	0.094	0.890	0.177	0.087	0.152

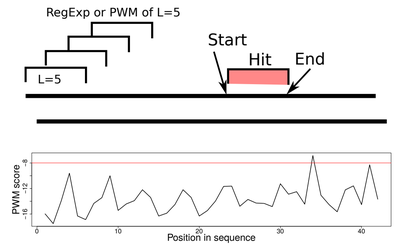


Matt allows the user to

1. **Learn PWMs** from aligned sequences
2. **Plot sequence logos** for PWMs
3. **Read-in PWMs** from output of other programs if given as table

Predicting hits of binding motifs in sequences

E.g., with human SF1 binding motif in 3' ends of introns

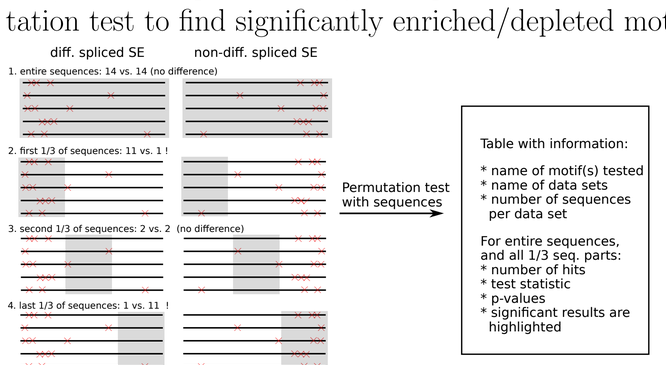


SEQ_ID	START	END	STRAND	HIT	LOGSCORE
Int01	1838	1844	-	TACTAAT	4.27
Int01	1471	1477	+	CACTAAA	5.02
Int02	94	100	+	AACTGAA	4.03
Int02	375	381	+	TACTAAA	4.39
Int03	1076	1082	+	AAATGAC	4.05
Int03	0	6	-	CACTCAC	4.85
Int04	1008	1014	+	CACTGAA	4.43
...	...	...	...	...	...

Matt **includes approx. 300 binding motifs** from CISBP-RNA (a database of RBPs, [3]) for automatic prediction and enrichment analysis.

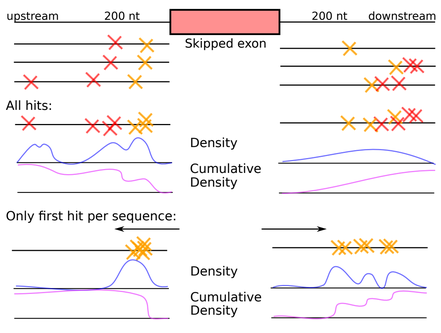
Enrichment analysis with hits of binding motifs

E.g., having extracted and defined two sequence sets, e.g. upstream sequences of diff./non-diff. spliced exons, Matt predicts hits of binding motifs and applies a permutation test to find significantly enriched/depleted motifs.



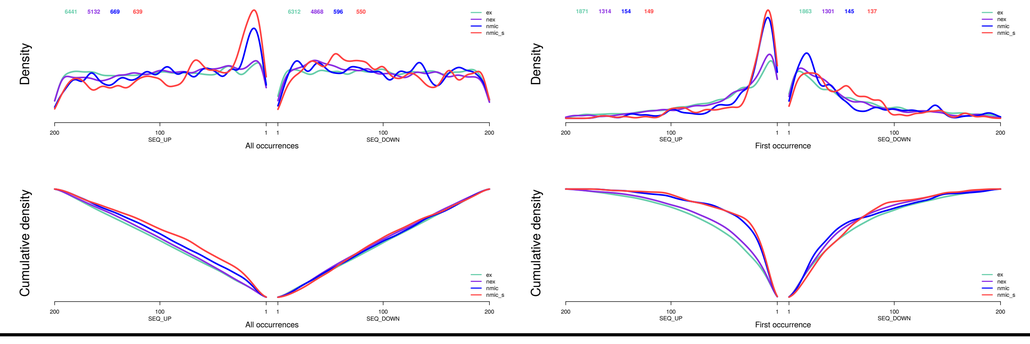
Applying permutation test to parts of sequences allows the **detection of positional bias**. Matt offers **automatic enrichment analysis** with all approx. **300 binding motifs** from CISBP-RNA ([3]).

Plotting distribution of hits of binding motifs



Matt predicts hits of binding motifs and plots positional densities for **all hits** and **only first hits**. Considering only first hits might show more clearly positional bias of hits.

E.g., TGC hits (bound by neural splicing regulator nSR100) in up-/downstream regions of neural micro-exons (mic, mic_s), neural exons (nex), constitutive exons (ex).



Affiliation

[1] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain

[2] Universitat Pompeu Fabra (UPF), Barcelona, Spain

References

1. Yeo, G. and Burge, CB. (2004). *J. Comp. Biol.*, **11**(2-3), 377–394.
2. Corvelo et. al, (2010). *PLoS Comput. Biol.*, **6**(11), e1001016.
3. Ray D. et al., (2013). *Nature* **499**(7457), 172–177.

Funding

This work has been supported by grants from European Research Council (ERC-StG-LS2-637591, ERC-AdG-MASCP-670146) and Spanish Ministry of Economy and Competitiveness (BFU2014-55076-P, BFU2014-55058-P). CRG is a Severo Ochoa Center of Excellence 2013-2017.